

SOFTWARE

Open Access



Informeasure: an R/bioconductor package for quantifying nonlinear dependence between variables in biological networks from an information theory perspective

Chu Pan^{1*} and Yanlin Chen²

*Correspondence:
chu.pan@hnu.edu.cn

¹ College of Computer Science
and Electronic Engineering,
Hunan University, Changsha,
Hunan, China

² School of Software, Henan
University of Engineering,
Zhengzhou, Henan, China

Abstract

Background: Using information measures to infer biological regulatory networks can capture nonlinear relationships between variables. However, it is computationally challenging, and there is a lack of convenient tools.

Results: We introduce Informeasure, an R package designed to quantify nonlinear dependencies in biological regulatory networks from an information theory perspective. This package compiles a comprehensive set of information measurements, including mutual information, conditional mutual information, interaction information, partial information decomposition, and part mutual information. Mutual information is used for bivariate network inference, while the other four estimators are dedicated to trivariate network analysis.

Conclusions: Informeasure is a turnkey solution, allowing users to utilize these information measures immediately upon installation. Informeasure is available as an R/Bioconductor package at <https://bioconductor.org/packages/Informeasure>.

Keywords: R package, Information measure, Nonlinear dependence, Biological regulatory network

Introduction

Quantifying the dependence between variables in biological networks is crucial, as understanding condition-responsive activations of biological processes relies on accurately constructing these networks [1]. Several computing criteria are used to quantify relationships either between paired variables or among multiple variables [2]. Correlation methods, such as Pearson correlation coefficient and partial correlation, are widely employed to assess the linear relationship between paired variables. In contrast, information measures can evaluate the dependence between two or more variables, offering significant advantages over correlation methods. These advantages include the ability to capture more general nonlinear associations and reflect dynamics between variables. Commonly used information measures include mutual information (MI), conditional



© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

$$\begin{aligned} \text{CMI}(X; Y|Z) &= \sum_{x \in X, y \in Y, z \in Z} p(x, y, z) \log \left(\frac{p(x, y|z)}{p(x|z)p(y|z)} \right) \\ &= H(X, Z) + H(Y, Z) - H(X, Y, Z) - H(Z) \end{aligned} \quad (3)$$

PMI is another measure that dedicates to quantifying the nonlinear direct dependence between two random variables given a third, particularly if any one variable has a potentially strong correlation with the third. For the discrete variables X , Y and Z , the formula for PMI is:

$$\text{PMI}(X; Y|Z) = \sum_{x \in X, y \in Y, z \in Z} p(x, y, z) \log \left(\frac{p(x, y|z)}{p^*(x|z)p^*(y|z)} \right) \quad (4)$$

where $p(x, y|z) = p^*(x|z) \cdot p(y|z)$ is the partial independence of discrete variable X and Y given Z . Here $p^*(x|z)$ and $p^*(y|z)$ are defined as:

$$p^*(x|z) = \sum_{y \in Y} p(x|z, y)p(y), \quad p^*(y|z) = \sum_{x \in X} p(y|z, x)p(x) \quad (5)$$

Another multivariate information measure is Π , which quantifies the amount information (synergy or redundancy) contained in a set of variables beyond any subset of those variables. The Π value can be either negative or positive. For three-variable cases, a positive interaction information value typically indicates synergistic or cooperative relationships among the first two variables given the third, meaning that the combined information from the variables provides more insight together than individually. In contrast, a negative interaction information value suggests redundant or suppressive interactions between the first two variables, where knowing the first two variables together yields less information than expected from their individual pairwise interactions [17].

The formula for Π is expressed as:

$$\Pi(X; Y; Z) = \text{MI}(X; Y) - \text{CMI}(X; Y|Z) \quad (6)$$

PID is another emerging information measure that decomposes the source information acting on a target into four parts: joint information (synergy), individual information (two unique components), and shared information (redundancy). For three discrete variables X , Y and, the formula for PID is:

$$\begin{aligned} \text{PID}(X; Y, Z) &= \text{Synergy}(Z; X, Y) + \text{Unique}_Y(Z; X) \\ &\quad + \text{Unique}_X(Z; Y) + \text{Redundancy}(Z; X, Y) \end{aligned} \quad (7)$$

Synergy represents the additional information about Z provided by X and Y together, not by each individual variable. Redundancy refers to the portion of information about Z provided by either variable X or Y alone. The unique contribution from X (or Y) is the part of information provided only by X (or Y).

In summary, MI is suitable for quantifying nonlinear dependence between two variables in biological networks, while CMI and PMI are ideal for trivariate network inference. The synergistic and redundant information decomposed from Π and PID can help biologists better decipher the cooperative and competitive relationships in more complex biological networks.

Implementation and main functions

Our toolkit is an extension of the entropy R package. We leveraged three distinct entropy estimators from this package. Our key innovation lies in extending these entropy estimators to five unique information measures, which can be applied directly to pairs or triples of variables, offering enhanced versatility in analyzing variable interactions.

The implementation of information measures typically involves first discretizing continuous variables into a count table, evaluating probabilities from the counts, estimating entropy based on the (joint) probability matrix, and finally calculating the information value associated between the variables. The package breaks the entire implementation process into three main parts: data discretization, probability/entropy estimation, and information estimation. This user-oriented implementation allows users to freely combine discretization methods, probability or entropy evaluators, and information metrics according to their specific requirements.

Two of the most common discretization methods are adopted in this package. The default method is a uniform width-based method, which divides the continuous data into N bins with equal width. The alternative is a uniform frequency-based approach, which divides the continuous data into N bins with an equal count. By default, both methods initialize the number of bins to the round-off value of the square root of the data size: \sqrt{N} .

In the probability estimation process, three types of probability estimators, referenced from the 'entropy' package, are available: the empirical estimator (default), the Dirichlet distribution estimator, and the shrinkage estimator. The Dirichlet distribution estimator includes four different distributions with different prior values:

- **method = "ML"**: maximum likelihood estimator, also referred to empirical probability,
- **method = "Jeffreys"**: Dirichlet distribution estimator with prior $a = 0.5$,
- **method = "Laplace"**: Dirichlet distribution estimator with prior $a = 1$,
- **method = "SG"**: Dirichlet distribution estimator with prior $a = \frac{1}{\text{length}(XY)}$, where XY is the joint count table for variables X and Y ,
- **method = "minimax"**: Dirichlet distribution estimator with prior $a = \sqrt{\frac{\sum(XY)}{\text{length}(XY)}}$,
- **method = "shrink"**: shrinkage estimator.

The most important functions in this package are the five different information measures, each ending with '.measure()' as the postfix. They are 'MI.measure()' for MI, 'CMI.measure()' for CMI, 'II.measure()' for II, 'PID.measure()' for PID and 'PMI.measure()' for PMI. Each function can be called with just a joint count table, but they also provide six probability estimation methods and three different base logarithmic calculations for users to choose from. All functions, except 'PID.measure()', return a numeric value representing the information measure between two variables or among three variables.

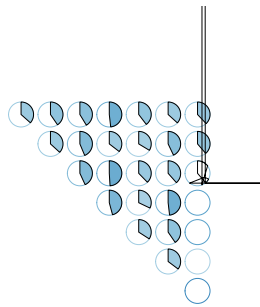
The 'PID.measure()' function returns a list that includes synergistic information, unique information from one source variable, unique information from the other source variable, redundant information, and the sum of the four parts of information.

Application in inferring different transcriptome regulatory network types

We consider estimating information measures from breast cancer expression profile data generated by the Cancer Genome Atlas (TCGA, <https://portal.gdc.cancer.gov>).

This sample data is stored in the 'extdata' directory and is used for various types of transcriptome regulatory network inferences, as shown in Fig. 1. All relevant details regarding the implementation of the following examples, including discretization and probability estimation methods, were thoroughly documented in the software vignette.

For two variables, we apply MI to identify the dependence between proteins in protein-protein interaction network inference. Figure 1A shows that the MI value of BRCA1 and BARD1 is the highest among all pairs, which is consistent with the evidence that these two proteins come from a complex and are highly dependent in nature.



In the three-variable case, we apply CMI and PMI for the triplet network inference, as both measures evaluate the influence of a third variable on the mutual information of two joint variables. Such characteristics make them suitable for ceRNA network inference. For example, in lncRNA-associated ceRNA triplets, the calculation pertains to the perturbation intensity on miRNA-mRNA by lncRNA. As shown in Fig. 1B, the PMI value is higher than the CMI value. This discrepancy arises because PMI accurately can evaluate the dependence between hsa-miR-26a-5p and PTEN, even though PTENP1 has a strong correlation with PTEN. In contrast, CMI tends to underestimate the result in this scenario.

For exploring the cooperative or competitive regulation mechanisms of two miRNAs on a common target mRNA, we apply II and PID to extract synergetic and/or redundant information between variables. As shown in Fig. 1C and D, the positive value of II and the high synergy value of PID indicate that the two miRNAs together provide most of the information for the target variable, suggesting a cooperative mechanism between the two miRNAs. The boxplot demonstrates that the expression patterns of hsa-miR-34a-5p and hsa-miR-34b-5p are similar yet distinct from the target MYC, which is consistent with the notion that the two miRNAs coordinately regulate the common target mRNA.

Discussion and conclusion

Our goal is to provide researchers with a comprehensive toolkit of information measures to address specific research purposes. The developed tool, Informeasure, is an R/Bioconductor package with well-documented functions and demonstration examples in the vignette, allowing users to easily access these information measures. In the current version, our primary focus is on applying information measures to two- and three-variable cases, although measures such as II and PID can potentially be extended to higher dimensions. However, to the best of our knowledge, identifying nonlinear dependence between two- and three-variable is currently the main concern. Therefore, we have chosen three variables as the largest network unit handled by this toolkit.

We recommend applying appropriate normalization methods specific to the data type before feeding the data into our toolkit. As demonstrated in our software vignette, RNA-seq data, which can have a highly skewed distribution, benefits from log₂ transformation to bring expression values to a comparable scale before discretization. For single cell RNA-seq data, we suggest using a global-scaling normalization method like “Log-Normalize” which is used by default in the Seurat package to handle sparse single cell data [18]. Additionally, we recommend using imputation strategies such as MAGIC or SAVER to address dropout issues for single cell RNA-seq data before applying our package [19, 20]. In the case of qPCR data, which often features a narrower range of expression values, log₂ transformation can still be beneficial depending on the dynamic range of the data.

In conclusion, we have implemented five information measures in this R package. A brief survey of information theory guided users in choosing appropriate measures for specific purposes. The illustrations successfully demonstrate the application of information measures for inferring various types of regulatory networks from expression profile data, with a primary focus on trivariate networks. We are convinced that Informeasure

can be widely serving as a valuable tool for facilitating the inference of condition-specific regulatory networks.

Availability and requirements

Project name: Informeasure

Project home page: <https://bioconductor.org/packages/release/bioc/html/Informeasure.html>

Operating system(s): Platform independent

Programming language: R

Other requirements: R \geq 4.0

License: None

Any restrictions to use by non-academics: None

Acknowledgements

The authors thank Dr. Junpeng Zhang, Mr. Nitesh Turaga, and Mr. Martin Morgan for their informative suggestions on writing the R package.

Author Contributions

CP conceptualization, writing-original software, supervision, funding acquisition, writing-original draft, project administration, writing-review and editing. YC writing-original software, writing-review and editing

Funding

This work was supported by National Natural Science Foundation of China [62102144 to C.P.]

Availability of data and materials

The RNA-seq dataset for breast cancer was downloaded from The Cancer Genome Atlas (TCGA, <https://portal.gdc.cancer.gov>)

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing of interests

The authors declare that they have no competing of interest.

Received: 9 August 2024 Accepted: 21 November 2024

Published online: 18 December 2024

References

- Hill SM, Heiser LM, Cokelaer T, Unger M, Nesser NK, Carlin DE, Zhang Y, Sokolov A, Paull EO, Wong CK, et al. Inferring causal molecular networks: empirical assessment through a community-based effort. *Nat Methods*. 2016;13(4):310–8.
- Maetschke SR, Madhamshettiwar PB, Davis MJ, Ragan MA. Supervised, semi-supervised and unsupervised inference of gene regulatory networks. *Brief Bioinform*. 2014;15(2):195–211.
- Wyner AD. A definition of conditional mutual information for arbitrary ensembles. *Inf Control*. 1978;38(1):51–9.
- McGill W. Multivariate information transmission. *Trans IRE Prof Group Inform Theory*. 1954;4(4):93–111.
- Williams PL, Beer RD. Nonnegative decomposition of multivariate information. 2010. arXiv preprint [arXiv:1004.2515](https://arxiv.org/abs/1004.2515).
- Zhao J, Zhou Y, Zhang X, Chen L. Part mutual information for quantifying direct associations in networks. *Proc Natl Acad Sci*. 2016;113(18):5130–5.
- Zhang X, Zhao X-M, He K, Lu L, Cao Y, Liu J, Hao J-K, Liu Z-P, Chen L. Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information. *Bioinformatics*. 2012;28(1):98–104.
- Zhang X, Liu K, Liu Z-P, Duval B, Richer J-M, Zhao X-M, Hao J-K, Chen L. Narromi: a noise and redundancy reduction technique improves accuracy of gene regulatory network inference. *Bioinformatics*. 2013;29(1):106–13.
- Sumazin P, Yang X, Chiu H-S, Chung W-J, Iyer A, Llobet-Navas D, Rajbhandari P, Bansal M, Guarnieri P, Silva J, et al. An extensive microRNA-mediated network of rna-rna interactions regulates established oncogenic pathways in glioblastoma. *Cell*. 2011;147(2):370–81.
- Chiu H-S, Llobet-Navas D, Yang X, Chung W-J, Ambesi-Impiombato A, Iyer A, Kim HR, Seviour EG, Luo Z, Sehgal V, et al. Cupid: simultaneous reconstruction of microRNA-target and cerna networks. *Genome Res*. 2015;25(2):257–67.

11. Salmena L, Poliseno L, Tay Y, Kats L, Pandolfi PP. A Cerna hypothesis: the Rosetta stone of a hidden RNA language? *Cell*. 2011;146(3):353–8.
12. Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Favera RD, Califano A. Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinform*. 2006;7:1–15.
13. Chan TE, Stumpf MP, Babbie AC. Gene regulatory network inference from single-cell data using multivariate information measures. *Cell Syst*. 2017;5(3):251–67.
14. Stumpf PS, Smith RC, Lenz M, Schuppert A, Müller F-J, Babbie A, Chan TE, Stumpf MP, Please CP, Howison SD, et al. Stem cell differentiation as a non-Markov stochastic process. *Cell Syst*. 2017;5(3):268–82.
15. Meyer PE, Lafitte F, Bontempi G. minet: Ar/bioconductor package for inferring large transcriptional networks using mutual information. *BMC Bioinform*. 2008;9:1–10.
16. Hausser J, Strimmer K. Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks. *J Mach Learn Res*. 2009;10(7).
17. Timme N, Alford W, Flecker B, Beggs JM. Synergy, redundancy, and multivariate information measures: an experimentalist's perspective. *J Comput Neurosci*. 2014;36:119–40.
18. Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, Lee MJ, Wilk AJ, Darby C, Zager M, et al. Integrated analysis of multimodal single-cell data. *Cell*. 2021;184(13):3573–87.
19. Van Dijk D, Sharma R, Nainys J, Yim K, Kathail P, Carr AJ, Burdziak C, Moon KR, Chaffer CL, Pattabiraman D, et al. Recovering gene interactions from single-cell data using data diffusion. *Cell*. 2018;174(3):716–29.
20. Huang M, Wang J, Torre E, Dueck H, Shafer S, Bonasio R, Murray JI, Raj A, Li M, Zhang NR. Saver: gene expression recovery for single-cell rna sequencing. *Nat Methods*. 2018;15(7):539–42.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.